Semantic features for context organization

Mário Antunes
Instituto de Telecomunicações
Universidade de Aveiro
Aveiro, Portugal
Email: mario.antunes@av.it.pt

Diogo Gomes
Instituto de Telecomunicações
Universidade de Aveiro
Aveiro, Portugal
Email: dgomes@av.it.pt

Rui Aguiar
Instituto de Telecomunicações
Universidade de Aveiro
Aveiro, Portugal
Email: ruilaa@av.it.pt

***Abstract*—In recent years the technological world has grown by incorporating billions of small sensing devices, collecting and sharing real-world information. As the number of such devices grows, it becomes increasingly difficult to manage all these new information sources. There is no uniform way to share, process and understand context information. In previous publications we discussed efficient ways to organize context information that is independent of structure and representation. However, our previous solution suffers from semantic sensitivity. In this paper we review semantic methods that can be used to minimize this issue, and propose an unsupervised semantic similarity solution that combines distributional profiles with public web services. Our solution was evaluated against Miller-Charles dataset, achieving a correlation of 0.6.**

***Keywords*—Internet of things, M2M, context information**

I. Introduction

Today the technological world is full of devices with sensing capabilities. Such concentration of computational capabilities is a direct consequence of the Internet of Things (IoT) and enables complex Machine-2-Machine (M2M) scenarios. These devices generate massive amounts of data, which are an untapped source of context information.

In Machine-to-Machine (M2M) scenarios, an entity's context can be used to provide added value: improve efficiency, optimize resources and detect anomalies. The following examples illustrate the importance of context information in M2M scenarios. Fusing data from several sensors makes it possible to predict a driver's ideal parking spot [1], [2]. Projects such as Pothole Patrol[3] and Nericell [4] use vehicular accelerations to monitor road conditions and detect potholes. Transport Information Monitoring Environment (TIME) project [5] combines data from mobile and fixed sensors in order to evaluate road congestion in real time.

These projects provide valuable insight about context information potential in advanced context-aware applications. However, many of these projects follow a vertical approach. This has hindered interoperability and the realisation of even more powerful IoT scenarios. Another important issue is the need felt for a new way to manage, store and process such diverse machine made context information; unconstrained and without limiting structures.

In previous publications we addressed some of these issues [6]–[8]. Such as the fact that devices/manufacturers share context information with a different structure, leading to information silos and low interoperability in M2M scenarios. One important objective of context representation research [9]–[11] is to standardize the process of sharing (with different platforms) and understanding context information.

We modelled context organization as an information retrieval problem. These systems commonly rely on the vector space model (VSM) [12] to compute the relevance ranking between documents and queries. However, this model has some drawbacks, the most relevant for our scenario is semantic sensitivity. Documents with similar context but with different vocabulary will not be associated, producing a false negative. This implies that context organization is highly depended on the document's vocabulary.

In this paper we explore semantic methods with the objective to minimize semantic sensitivity. By using semantic methods it is possible to organize, extract and cluster information based on concepts and not on sub-strings nor regular expressions. Apart from context-aware applications, several other areas benefit from semantic based context organization. Without loss of generality let us assume that given a set of M2M devices we are able to autonomously build a concept tree. A concept tree is a tree like structure where concepts are organized from the broader to the most specific. The previous structure can be used to optimize information retrieval system for M2M scenarios. Given a query it is possible to determine the most relevant topic by traversing the concept tree. Machine learning algorithms, specially pattern matching algorithms, can use a concept tree to prune large portions of lesser relevant information. Finally, these aspects could provide a decisive contribution towards the exploration of name-based information centric network architectures in IoT environments[13]. Namely, the application of inference mechanisms into the content-reaching operations of the networking fabric itself can be used to have the network better mimic the complex relationships between devices (e.g., sensors, actuators), their generated content (e.g., temperature values with different units) and its dissemination towards interested entities.

The remainder of the paper is organized as follows. In Section II we define common characteristics of M2M context information and analyse how this can be organized. We review the vector space model and semantic methods in Section III and Section IV respectively. Section V contains implementation details of our prototype. The results of our

evaluation are in Section VI. Finally, the discussion and conclusions are presented in Section VII.

II. Context organization model

Context information is an enabler for further data analysis, potentially exploring the integration of an increasing number of information sources. The common definitions of context information [14], [15] are so broad that any information related to an entity can be considered context information. These definitions also do not provide any insight about the structure of context information. As previously mentioned, currently no uniform way to share/ manage vast amounts of M2M information. M2M devices commonly share information in textual format. From now on we will refer to a unique piece of context information as a document, and an entity that produces context information as a source.

Context information can be organized using two different approaches: top-down and bottom-up [16]–[18]. Top-down characterization requires the definition of classes and their relations *a priori* (similar to taxonomies and ontologies). Due to the diversity and vast amount of M2M devices it is very difficult to define and maintain such classes and the relations between them. On the other hand, bottom-up characterization is massively dimensional, and there is no global consistency imposed by current practice. Each individual piece of information is divided into features. *A posteriori* some metrics and learning algorithms can be used to find patterns and relations.

Our first solution [6] used a bottom-up 1-dimension model (see Figure 1). Each document is identified with a unique key, without any regards by its source.

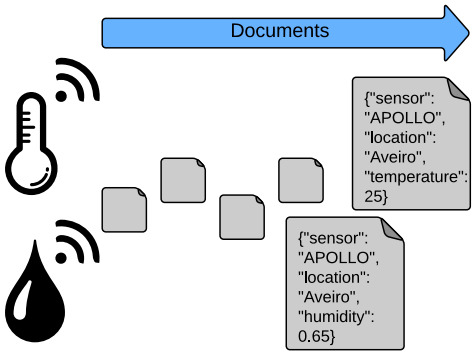

Fig. 1. Representation of a 1-dimension model.

This model has some drawbacks: poor scalability and semantic extraction. It is important that we analyse the typical traffic behaviour in M2M scenarios. For instance the majority of sensors send information periodically or when a specific event is detected. As such context information is better modelled as continuous document streams than as a set of independent documents.

Another issue that we must be aware is that the majority of the documents are represented in semi-structured format (e.g. XML, YAML, JSON, BSON). Most common semi-structured representations can be mapped into an entity-attribute-value (EAV) model [19]. The source is the **Entity**, and each document is a set of pairs **Attribute**/**Value**. The semantic value of a document is in the **Attributes**, while the **Values** are variables that change over time.

Taking these features into account we proposed a d-dimension (see Figure 2). The first dimension is always the source and the remaining $d-1$ dimensions are used to filter data from a specific source [7], [8]. This model uses a bottom-up approach to organize the stream's semantic portion and the remaining $d-1$ can be explored as an OLAP cube.

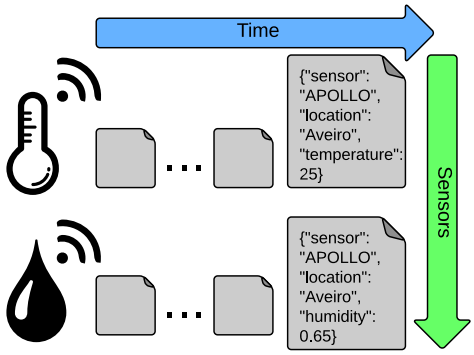

Fig. 2. Representation of a 2-dimension model. The first and second dimensions are source and time respectively.

In this paper we continue to expand this idea. M2M devices share a vast diversity of information. However, we can classify the information into two distinct classes: semantically rich and poor. In order to better understand these concepts let us consider the following example. A sensor node in a green house measures 6 effects: air and soil temperature, air and soil humidity, CO2 and leaf wetness. The node can periodically share the measurements individually or grouped in a single file. Each document shared in the first option are semantically poor. Based on the semantic value of its **attributes** it is quite difficult to associate the green house concept with each stream individually. By contrast, the single document with all the **attributes** is closer to the green house concept.

We can improve our organization model based on this observation. Through semantic methods it is possible to learn/extract higher level concepts from semantically rich documents. The end game is to propagate these concepts to the other stream based on similarity. The similarity between streams is calculated based on the remaining $d-1$ dimensions and the stream itself.

III. Vector space model

As previously mentioned, the best option to organized context information is through a bottom-up approach. Although M2M streams are not usually tagged by users, we can decompose the stream's semantic portion into discriminative concepts. A concept is a sequence of one or more words that provide a compact representation of a document's content. Ideally, concepts represent in condensed form the essential content of a document.

In our view organization models can be achieved through a bottom-up characterization considering it an information

retrieval problem. Organizing documents based on its content is one of the major objectives of information retrieval research: information retrieval informs on the existence (or non-existence) and whereabouts of documents related with user's query (similar to a web search engine). These systems commonly rely on the vector space model (VSM) to compute the relevance ranking between documents and queries. However, this model relies on sub strings to determine the relevance between terms and documents. In this section we review the vector space model and point out the most relevant drawbacks for M2M scenarios.

Vector space model is an algebraic model for representing documents and queries as vectors of terms. It is extensively used in information filtering, information retrieval, indexing and relevancy rankings. Each dimension corresponds to a separate term, typically weighted by tf-idf [20]. The relevance between documents and queries is computed with cosine similarity.

As previous mention the vector space model has some drawbacks, the most relevant is semantic sensitivity. Documents with similar context but with different vocabulary will not be associated, producing a false negative. This is specially worrisome for M2M scenarios, since there is no uniform way to share context information. Context organization becomes highly depended on the document's vocabulary. In Section IV we review well known semantic methods, which can be used to minimize this issue.

IV. Semantic methods

Some of the most popular semantic methods are based on latent analysis [21]–[23]. These well known methods analyse the co-occurrences of terms in a corpus of documents in order to find hidden/latent variables, regarded as topics or concepts. Since the number of concepts is usually greatly inferior to the number of words and it is not necessary to know the document categories/classes,these methods are thus unsupervised dimensionality reduction techniques. Common applications include information retrieval, document classification and collaborative filtering.

Without loss of generality let us analyse latent semantic analysis (LSA). LSA assumes that words that are close in meaning will occur in similar pieces of text. A matrix containing word counts per passage (rows and columns represent unique words and passages respectively) is constructed from a large corpus. The matrix is reduced using a factorization, called singular value decomposition (SVD), while preserving the similarity structure among columns. The language-theoretical interpretation of the result of the analysis is that LSA vectors approximate the meaning of a word as its average effect on the meaning of passages in which it occurs, and reciprocally approximates the meaning of passages as the average of the meaning of their words.

These methods work well in large corpus with a vast vocabulary. Although the amount of information associated with M2M scenarios is large, its vocabulary is rather poor. The majority of information is generated automatically by devices and share in a semi-structure format. Another disadvantage, is the fact that latent variables represent concepts that might be difficult to interpret. This leads to results which can be justified on the mathematical level, but have no interpretable meaning in natural language.

Other popular semantic methods are estimating distance between two units of language. Semantic distance is a measure of how close or distant two units of language are, in terms of their meaning. For example, the nouns *banana* and *fruit* are closer in meaning than the nouns *banana* and *car*.

Two classes of automatic semantic distance measures exist. **Lexical-resource-based measures of concept-distance** rely on the structure of a knowledge source, such as WordNet[24], to determine the distance between two concepts. In the WordNet database nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. WordNet superficially resembles a thesaurus, in that it groups words together based on their meanings. However, there are some important distinctions. First, WordNet interlinks not just word forms (strings of letters) but specific senses of words. As a result, words that are found in proximity to one another in the network are semantically disambiguated. Second, WordNet labels the semantic relations among words, whereas the groupings of words in a thesaurus does not follow any explicit pattern other than meaning similarity. Several authors proposed semantic measures based on WordNet [25]–[27].

However, creating and maintaining lexical databases is a tedious task that requires human experts. Further, updating a resource is again expensive and there is usually a lag between the current state of language usage/comprehension and the lexical resource representing it. For example, due to funding and staffing issues the WordNet project is no longer accepting comments and suggestions[1]. Although M2M information has limited vocabulary, usually consists of very specific terms associated with the technology world. The lexical database may not contain these terms or the correct associations.

Distributional measures of word-distance rely on the **distributional hypothesis**, which states that words that occur in similar contexts tend to be semantically close [28], [29]. Many distributional approaches represent the sets of contexts of the target words as points in multidimensional co-occurrence space. Some metrics can be used to measure distributional distance, such as cosine and α-skew divergence [30] among others.

These methods work well and do not require a lexical database. The distributional profile can be used to enrich the vocabulary in present M2M scenarios. However, these methods require a large corpus, which is a disadvantage. Due to the poor vocabulary present in M2M scenarios, the corpus made up from the information shared by M2M devices is not suitable to learn distributional profiles. Creating and maintaining a large corpus for M2M scenarios is a time intensive task that requires human experts. It has the same exact problems of creating and maintaining a large lexical database. The vast amount and diversity of

[1] http://wordnet.princeton.edu/wordnet/

information and the poor vocabulary represent additional difficulties.

In this paper we explore the idea of using external public services as a replacement for a large corpus. Conventional search engines provide access to vast amounts of documents with rich vocabulary. On-line encyclopedias such as Wikipedia[2], Scholarpedia[3] and Citizendium[4] provide access to high quality definitions and definitions. Finally, web thesaurus can be used to optimize distributional profiles by reducing dimensions that are synonyms.

V. Implementation

In this section we discuss important details about our solution. Given a word our solution prototype uses a web search engine to extract its distributional profile. These profiles can be evaluated with the cosine similarity. Our prototype is divided into 3 different component as depicted in Figure 3. Currently our prototype only works with English vocabulary. All the components were prototyped in Java.

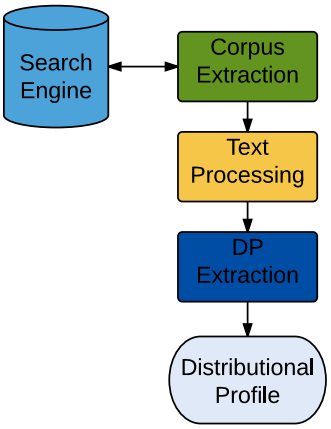

Fig. 3. Proposed DP extraction system's architecture.

The first component (corpus extraction) bridges our solution with web search engines. Currently it uses Bing Search API[5], although other search engines can be instantiated and used. The basic function of this component is to extract a corpus, from a web search engine, that is associated with a specific word. Two different methods to extract a corpus were implemented: based on full pages, and based on snippets. The first method downloads and returns the content of each page indicated by the search engine. On the other hand, the second method only returns the snippets provided by the search engine.

The second component (text processing) implements a pre-processing pipeline that processes and cleans the corpus. The various spaces of the pipeline are depicted in Figure 4.

The sentence segmentation and tokenizer phases divide the corpus into sentences and tokens respectively. After,

[2]https://www.wikipedia.org/
[3]http://www.scholarpedia.org/
[4]http://en.citizendium.org/
[5]https://datamarket.azure.com/dataset/bing/searchweb

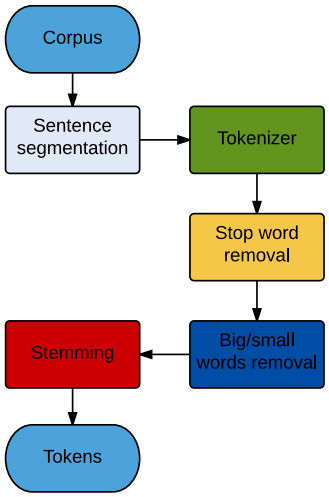

Fig. 4. Text processing pipeline.

tokens present in the stop word list are removed. Stop words are deemed irrelevant because they occur frequently in the language and provide little information about any topic. In our prototype we used the MySQL stop word list[6]. For the exact same reason we also remove tokens that are too big or small. Any token with less than 3 or more than 15 (6 is average word length in English) characters were removed from the pipeline. Finally each token is stemmed using the Porter stemming algorithm[31].

The final component (DP extraction) analyses the output of the pipeline and extracts the distributional profile of the word. The profile is a vector containing the neighbourhood of the specified word. Term frequency is used as to weight the relevance of each neighbour.

VI. Performance evaluation

We evaluate the proposed method against Miller-Charles dataset [32], a dataset of 30 word-pairs rated by a group of 38 human subjects. Currently there is no word similarity database specific for M2M scenarios. Since we do not have the technical capacity to develop a specific dataset for M2M scenarios, in this first work we used a well known general proposed dataset. We intend to address this issue in future publications.

The word pairs are rated on a scale from 0 (no similarity) to 4 (perfect synonymy). We analyse the behaviour of the proposed measure with the number of web pages/snippets and the size of the word neighbourhood.

Pearson correlation was used to evaluate our distance measure against the ground truth. Correlation between sets of data is a measure of how well they are related. The correlation r can range from -1 to 1. An r of -1 indicates a perfect negative linear relationship between variables, an r of 0 indicates no linear relationship between variables, finally and an r of 1 indicates a perfect positive

[6]https://dev.mysql.com/doc/refman/5.1/en/fulltext-stopwords.html

linear relationship between variables. In short, the highest correlation indicates the most accurate solution.

The experimental results are presented in Figure 5 and Figure 6. The distributional profiles were extracted from web pages and snippets respectively.

From Figure 5 it is apparent that the first web pages achieved the highest correlation, independently of the neighbourhood distance. The web pages returned by the search engine are ranked based on topic relevance. From the first pages it is possible to extract the most relevant distributional profile dimensions. Extracting a profile from more web pages adds several low relevant dimensions. These low relevant dimensions add up and decrease accuracy. We also see that the correlation slowly increases as the number of web pages increases, although never reaching the performance obtained only with the first pages. As more web pages are used to extract a profile, the weight of the low relevant dimensions decreases, minimizing their impact on performance. It is worth mentioning that the best results were achieved with a neighbourhood distance of 3.

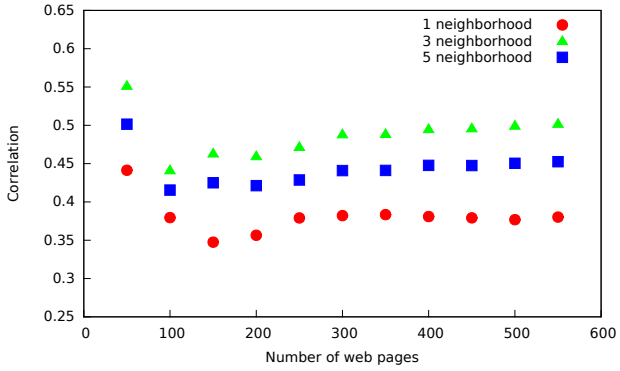

Fig. 5. Estimation of the index size.

Distributional profiles extracted from snippets achieved a higher accuracy overall, see Figure 6. For neighbourhood distance of 1 the first snippets achieved the highest correlation. Similarly, to the web pages experiment, the most relevant dimensions are present on the first snippets. Extracting a profile from more snippets only adds low relevant dimensions. On the other hand, neighbourhood distances of 3 and 5 achieved the highest accuracy with 250 snippets, after that point the accuracy decreases monotonically. It is apparent that higher neighbourhood distances achieved the best accuracy with snippets.

VII. Discussion and Conclusions

The number of sensing devices is increasing at a steady step. Each one of them generates massive amounts of information. However, each device/manufactures share context information with different structure, hindering interoperability in M2M scenarios. We proposed methods to organize context information that are independent from their representation and structure. Within this paper we explore semantic methods to improve our organizational model.

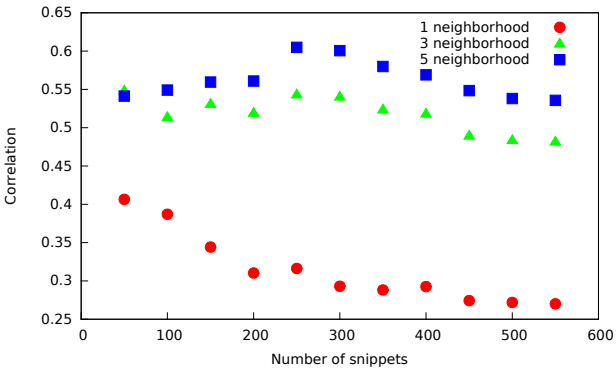

Fig. 6. Estimation of the index size.

We described our organization models and our previous approach to deal with the semantic portion of context information. This approach, based on Vector Space Model, suffers from semantic sensitivity. In order to minimize this drawback, we reviewed several well known semantic methods and evaluated their usability in M2M scenarios.

We proposed a semantic similarity solution that combines distributional profiles with public web services. Our solution was evaluated against Miller-Charles dataset [32], achieving a correlation of 0.6. There is room for improvement, on-line thesaurus and dimension reduction algorithms are possible options to optimize the profile. Other term weighting functions can be used, such as tf-idf [20]. Nevertheless, our unsupervised solution was able to learn distributional profiles from the web, achieving a relative high accuracy. The main advantages of our solution are: does not require a specific corpus, the profile can be used to enhance the poor vocabulary present is most M2M scenarios, and the profile's dimensions can be interpreted by human users.

Acknowledgement

This work has been partially funded by project Cloud Thinking (CENTRO-07-ST24-FEDER-002031), co-funded by QREN, "Mais Centro" program, under research grant SFRH/BD/94270/2013.

References

[1] T. Rajabioun, B. Foster, and P. Ioannou, "Intelligent parking assist," in *Control Automation (MED), 2013 21st Mediterranean Conference on*, June 2013, pp. 1156–1161.

[2] J. K. Suhr and H. G. Jung, "Sensor fusion-based vacant parking slot detection and tracking," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 15, no. 1, pp. 21–36, February 2014.

[3] J. Eriksson, L. Girod, B. Hull, R. Newton, S. Madden, and H. Balakrishnan, "The pothole patrol: Using a mobile sensor network for road surface monitoring," in *Proceedings of the 6th International Conference on Mobile Systems,Applications, and Services*, 2008, pp. 29–39.

[4] P. Mohan, V. N. Padmanabhan, and R. Ramjee, "Nericell: rich monitoring of road and traffic conditions using mobile smartphones," in *Proc. of the 6th ACM conference on Embedded network sensor systems*, 2008, pp. 323–336.

[5] J. Bacon, A. Bejan, A. Beresford, D. Evans, R. Gibbens, and K. Moody, "Using real-time road traffic data to evaluate congestion," in *Dependable and Historic Computing*, ser. Lecture Notes in Computer Science, C. Jones and J. Lloyd, Eds. Springer Berlin Heidelberg, 2011, vol. 6875, pp. 93–117.

[6] M. Antunes, D. Gomes, and R. Aguiar, "Context storage for m2m scenarios," in *Proc. ICC 2014*, 2014.

[7] ——, "Scalable semantic aware context storage," in *Future Internet of Things and Cloud (FiCloud), 2014 International Conference on*, Aug 2014, pp. 152–158.

[8] ——, "Semantic-based publish/subscribe for m2m," in *Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2014 International Conference on*, Oct 2014, pp. 256–263.

[9] R. M. Turner, "A model of explicit context representation and use for intelligent agents," in *Modeling and Using Context*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 1999, vol. 1688, pp. 375–388.

[10] T. Strang, C. Linnhoff-Popien, and K. Frank, "Cool: A context ontology language to enable contextual interoperability," in *Distributed Applications and Interoperable Systems*, ser. Lecture Notes in Computer Science, J.-B. Stefani, I. Demeure, and D. Hagimont, Eds. Springer Berlin Heidelberg, 2003, vol. 2893, pp. 236–247.

[11] O. Lassila and D. Khushraj, "Contextualizing applications via semantic middleware," in *Mobile and Ubiquitous Systems: Networking and Services, 2005. MobiQuitous 2005. The Second Annual International Conference on*, July 2005, pp. 183–189.

[12] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Communications ACM*, vol. 18, no. 11, pp. 613–620, November 1975.

[13] J. Quevedo, D. Corujo, and R. Aguiar, "A case for icn usage in iot environments," in *Global Communications Conference (GLOBECOM), 2014 IEEE*, December 2014, pp. 2770–2775.

[14] G. D. Abowd, A. K. Dey, P. J. Brown, N. Davies, M. Smith, and P. Steggles, "Towards a better understanding of context and context-awareness," in *Proc. of the 1st international symposium on Handheld and Ubiquitous Computing*, 1999, pp. 304–307.

[15] T. Winograd, "Architectures for context," *Hum.-Comput. Interact.*, vol. 16, no. 2, pp. 401–419, December 2001.

[16] C. Shirky, "Ontology is overrated: Categories, links, and tags," http://shirky.com/writings/ontology_overrated.html, May 2005, accessed: 22-07-2013.

[17] G. Avram, "At the crossroads of knowledge management and social software," *Electronic Journal of Knowledge Management*, vol. 4, no. 1, pp. 1–10, January 2006.

[18] T. Gruber, "Ontology of folksonomy: A mash-up of apples and oranges," *International Journal on Semantic Web and Information Systems*, vol. 3, no. 2, pp. 1–11, 2007.

[19] P. M. Nadkarni, L. Marenco, R. Chen, E. Skoufos, G. Shepherd, and P. Miller, "Organization of heterogeneous scientific data using the eav/cr representation," *Journal of the American Medical Informatics Association*, vol. 6, no. 6, pp. 478–493, 1999.

[20] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, vol. 28, pp. 11–21, 1972.

[21] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[22] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '99, 1999, pp. 50–57.

[23] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, March 2003.

[24] G. A. Miller, "Wordnet: A lexical database for english," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, November 1995.

[25] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*, ser. ACL '94, 1994, pp. 133–138.

[26] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, ser. IJCAI'95, 1995.

[27] S. Banerjee and T. Pedersen, "An adapted lesk algorithm for word sense disambiguation using wordnet," in *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, ser. CICLing '02, 2002, pp. 136–145.

[28] J. Firth, "A synopsis of linguistic theory 1935-55," *Studies in Linguistic Analysis*, 1957.

[29] Z. Harris, *Mathematical Structures of Language*. John Wiley and Son, 1968.

[30] L. Lee, "On the effectiveness of the skew divergence for statistical language analysis," in *Artificial Intelligence and Statistics*, 2001, pp. 65–72.

[31] M. F. Porter, "Readings in information retrieval," K. Sparck Jones and P. Willett, Eds. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997, ch. An Algorithm for Suffix Stripping, pp. 313–316.

[32] G. A. Miller and W. G. Charles, "Contextual correlates of semantic similarity," *Language and Cognitive Processes*, vol. 6, no. 1, pp. 1–28, 1991.